tvst

Article

Predicting Glaucoma Progression Requiring Surgery Using Clinical Free-Text Notes and Transfer Learning With Transformers

Wendeng Hu[1] **and Sophia Y. Wang**[1]

[1] Byers Eye Institute, Department of Ophthalmology, Stanford University School of Medicine, Palo Alto, CA, USA

Correspondence: Sophia Y. Wang, Byers Eye Institute, Department of Ophthalmology, Stanford University School of Medicine, 2452 Watson Court, Palo Alto, CA 94304, USA.
e-mail: sywang@stanford.edu

Received: October 11, 2021
Accepted: March 2, 2022
Published: March 30, 2022

Keywords: glaucoma; free-text clinical notes; machine learning; natural language processing; BERT; transfer learning; deep learning

Citation: Hu W, Wang SY. Predicting glaucoma progression requiring surgery using clinical free-text notes and transfer learning with transformers. Transl Vis Sci Technol. 2022;11(3):37, https://doi.org/10.1167/tvst.11.3.37

Purpose: We evaluated the use of massive transformer-based language models to predict glaucoma progression requiring surgery using ophthalmology clinical notes from electronic health records (EHRs).

Methods: Ophthalmology clinical notes for 4512 glaucoma patients at a single center from 2008 to 2020 were identified from the EHRs. Four different pre-trained Bidirectional Encoder Representations from Transformers (BERT)-based models were fine-tuned on ophthalmology clinical notes from the patients' first 120 days of follow-up for the task of predicting which patients would require glaucoma surgery. Models were evaluated with standard metrics, including area under the receiver operating characteristic curve (AUROC) and F1 score.

Results: Of the patients, 748 progressed to require glaucoma surgery (16.6%). The original BERT model had the highest AUROC (73.4%; F1 = 45.0%) for identifying these patients, followed by RoBERTa, with an AUROC of 72.4% (F1 = 44.7%); DistilBERT, with an AUROC of 70.2% (F1 = 42.5%); and BioBERT, with an AUROC of 70.1% (F1 = 41.7%). All models had higher F1 scores than an ophthalmologist's review of clinical notes (F1 = 29.9%).

Conclusions: Using transfer learning with massively pre-trained BERT-based models is a natural language processing approach that can access the wealth of clinical information stored within ophthalmology clinical notes to predict the progression of glaucoma. Future work to improve model performance can focus on integrating structured or imaging data or further tailoring the BERT models to ophthalmology domain–specific text.

Translational Relevance: Predictive models can provide the basis for clinical decision support tools to aid clinicians in identifying high- or low-risk patients to maximally tailor glaucoma treatments.

Introduction

Glaucoma is a leading cause of irreversible blindness worldwide. The estimated prevalence of glaucoma is rising, from an estimated 76 million in 2020 to 111.8 million in 2040.[1] For many patients, glaucoma is a slowly progressive disease that, with routine treatments, can be stable over long periods.[2] However, it can be difficult to identify which subset of glaucoma patients will progress to require invasive surgery, because each patient is a complex combination of clinical factors, such as glaucoma subtype,[3–5] glaucoma medication usage patterns,[3,5] surgical history,[3–6] and intraocular pressure (IOP),[5] among others. Prospectively distinguishing between patients with likely progressive or stable disease could enable more aggressive interventions or relax the burden of follow-up and testing in appropriate populations.

Predictive models to predict glaucoma progression have typically relied on testing data, such as retinal nerve fiber layer optical coherence tomography or visual fields (VFs). However, it has been a challenge to incorporate a patient's clinical history, which typically resides within the patient health record, into these predictive models. The adoption of electronic health records (EHRs) has presented an opportunity to develop machine learning and deep-learning models

based on these data to predict glaucoma progression. Structured EHR data, such as demographics and billing codes, are the easiest data to access within the EHR and have been shown to provide some predictive value for the progression of glaucoma.[7] However, granular clinical data such as patient presenting symptoms, medical and surgical history, and examination findings are difficult to extract and integrate into predictive models, as these data typically reside within free-text clinical notes in unstructured formats. It is unclear what is the best way to integrate information from free text into models, especially in ophthalmology where this field of research is nascent.

Recent advances in natural language processing (NLP) with deep learning have enabled initial studies integrating the free-text clinical data into prediction models in general biomedical domains. Several studies integrate the clinical data from the free-text notes with convolutional neural networks (CNNs)[8,9] and/or word embeddings trained with a Continuous Bag of Words model.[8] Bidirectional Encoder Representations from Transformers (BERT), a deep-learning language model with transformer-based neural architecture, enabled major breakthroughs in NLP with state-of-the-art results in many general language tasks, including text classification.[10–12] Studies from a variety of medical fields have applied BERT to clinical notes to perform prediction and classification, such as identifying cartilage lesions from radiology reports, predicting unplanned readmissions following arthroplasty, and others.[13,14] Thus, we hypothesized that applying BERT-based models to ophthalmology clinical notes may be a promising method for predicting which glaucoma patients will progress to surgery.

In this paper, we compare the performance of various BERT-based models with regard to the use of clinical free-text progress notes to predict glaucoma progression to surgery. Four different widely accepted BERT-based models—original BERT, BioBERT, RoBERTa, and DistilBERT—were selected. We demonstrated that BERT-based models can provide some predictive value for the progression of glaucoma based on the clinical free-text notes, representing the first efforts demonstrating the use of BERT-based models for ophthalmology clinical text.

Methods

Data Source, Study Cohort, and Population Characteristics

We identified from the Stanford Clinical Data Warehouse[15] unique adult patients from 2009 to 2018 who underwent incisional glaucoma surgery (Current Procedural Terminology codes 66150, 66155, 66160,66165, 66170, 66172, 66174, 66175, 66179, 66180, 66183, 66184, 66185, 67250, 67255, 0191T, 0376T, 0474T, 0253T, 0449T, 0450T, 0192T, 65820, 65850, 66700, 66710, 66711, 66720, 66740, 66625, and 66540) or who had two or more instances of a glaucoma diagnosis but did not undergo glaucoma surgery (International Classification of Disease [ICD]-10 codes H40- [except H40.0-], H42-, Q150-, and their ICD-9 equivalents). Surgical patients must have had at least 120 days of baseline follow-up prior to surgery. Non-surgical patients must have had at least 120 days of follow-up. In all, there were 748 surgical patients and 3764 nonsurgical patients. A flow diagram illustrating the cohort creation is presented in Figure 1. Population characteristics, including demographic information and baseline visual acuity and intraocular pressure, were summarized for the cohort using structured data available from the EHR,[16] with proportions for categorical variables and means and standard deviations for continuous variables. Research was approved by the Stanford Institutional Review Board and adhered to the tenets of the Declaration of Helsinki.

Clinical Notes Dataset

We identified the first three clinical progress notes from within the first 120 days of follow-up and combined them into a single document for each patient. This period was chosen because new patients to our institution frequently return on separate visits to complete baseline testing such as imaging and visual fields, and it is common to take approximately 3 months to complete the testing. The train, validation, and test datasets were randomly split with an approximate ratio of 8:1:1. The training set contained 601 surgical patients and 3011 nonsurgical patients. The validation dataset contained 63 surgical patients and 337 nonsurgical patients. The test dataset contained 84 surgical patients and 416 nonsurgical patients.

Modeling Approach

Overview

BERT-based language models are trained in an unsupervised manner over large corpora of text in order to generate representations of the words that reflect their use and meaning. The original BERT model has undergone a variety of different refinements, having been pre-trained on different types of corpora or with slightly different parameters, as summarized below. We evaluated four pre-trained BERT-based

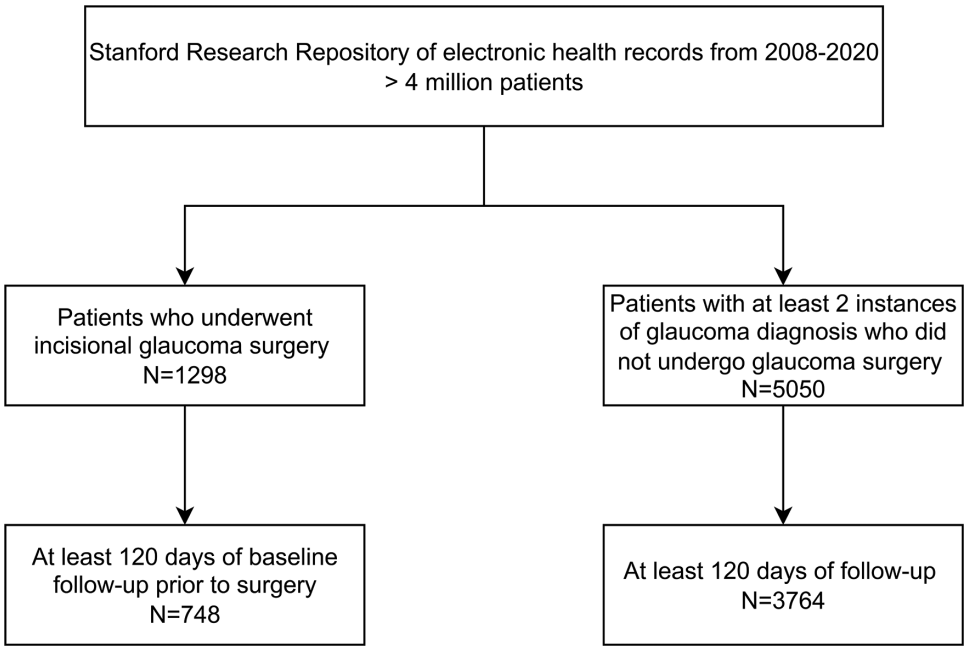

Figure 1. Flow diagram of patient identification and cohort creation.

models from the *huggingface* platform[17] for our task of predicting glaucoma progression requiring surgery using clinical free-text notes:

- BERT—BERT is short for Bidirectional Encoder Representations from Transformers, and its architecture relies on the attention mechanism to learn the inner representation of the language. It was pre-trained on a general English language corpus (BookCorpus and English Wikipedia), with masked language modeling and next sentence prediction tasks.[10] We applied *bert-base-uncased* (12 layers, 768 hidden states, 12 heads, 110 million parameters) to represent the key information from the clinical notes.
- BioBERT—BERT was trained on a general corpus that does not include domain-specific vocabulary. However, domain-specific terms contain key information in the biomedical text. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) was designed to improve our understanding of the biomedical text by pre-training on the biomedical domain corpus (PubMed abstracts and PubMed Central full-text articles) with BERT weights initialization.[18] *BioBERT-Base v1.1* (+PubMed 1M) was selected because of its reported excellent performance.
- RoBERTa—To improve the performance of BERT, a robustly optimized BERT approach (RoBERTa) was introduced by following the same training process as BERT with slight adjustments.[19] RoBERTa was trained for a longer time on dynamically masked longer sequences with larger batch sizes. We applied a *roberta-base* (12 layers, 768 hidden states, 12 heads, 125 million parameters) to predict the progression of glaucoma requiring surgery.
- DistilBERT—The BERT-based model is large and requires significant computational resources to train and perform inference with, limiting its application in practice. DistilBERT was proposed to compress the model size and maintain the same model performance.[20] More specifically, DistilBERT was trained on the same corpus as BERT to mimic the behavior of the BERT base model by returning the same probabilities and generating the same hidden states. We used the *distilbert-base-uncased* (6 layers, 768 hidden states, 12 heads, 66 million parameters) to encode the clinical text.

Data Preprocessing

Clinical notes were preprocessed to remove stopwords (a, all, also, an, and, are, as, at, be, been, by, for, from, had, has, have, in, is, it, may, of, on, or, our, than, that, the, there, these, this, to, was, we, were, which, who, with). All letters were processed to be lowercase. Because RoBERTa and BioBERT were pre-trained based on a cased vocabulary, in a sensitivity analysis we also fine-tuned these models on cased ophthalmology clinical text, with similar results. BERT-based models take as inputs each individual

word ("token") and break up rare words into multiple subwords to reduce the overall model vocabulary size. Subword tokenization algorithms were applied to our text to prepare it for input into our BERT-based models. After subword tokenization, clinical notes were truncated or padded to the required 512 token input length, which is the maximum length BERT can accept. BERT was designed as a fixed-length model with a maximum of 512 tokens, including the starting token, [CLS], and ending token, [SEP].

Model Training

All BERT-based models have similar model architectures and were fine-tuned following a previously described protocol.[10] Briefly, we removed the pretraining head of the model and added a randomly initialized linear classifier that was passed through a softmax function to obtain the probability of progression to glaucoma surgery. All parameters of the models were fine-tuned on our training dataset. Given that most glaucoma patients do not experience progression to surgery and to address class imbalance and prevent the model from learning a trivial classifier (i.e., predicting all patients will have no progression), we used a weighted cross-entropy loss function, which forces the model to pay more attention to predicting the minority class (progression to surgery) correctly. The weight was defined as the ratio of negative to positive labels, which was approximately 5. To perform hyperparameter tuning, we applied random search with early stopping to obtain the best possible combinations of hyperparameters for each model, as determined by model performance on the validation set. The hyperparameters that we tuned were learning rate, number of epochs, batch size, gradient accumulation, warm-up steps, and dropout rate. A summary of the final hyperparameters used is provided in Supplemental Table S1. All code for model training is publicly available.[21]

Evaluation

The F1 score (harmonic mean of precision and recall), sensitivity (recall), specificity, positive predictive value (precision), and negative predictive value were used to measure the performance of the models in predicting the progression of glaucoma requiring surgery on a held-out test set. The output of each model is a probability score that must be converted to binary predictions based on a selected probability threshold, which impacts those evaluation metrics. We tuned the probability threshold on the validation dataset to get the optimal F1 score. The selected threshold was applied to calculate other metrics above. To measure model performance independent of specific classification thresholds, the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) were selected. To provide a baseline for human-level prediction performance, a glaucoma specialist (SYW) reviewed the charts of a sample of 300 patients from the held-out test set and performed clinical predictions on whether they would progress to requiring surgery based on their clinical notes.

Interpretability

Because deep-learning neural networks tend to be black boxes for which it is difficult to explain the logic behind the prediction, we also qualitatively evaluated our fine-tuned BERT model by performing explainability studies based on the self-attention mechanisms of the BERT models,[12] similar to previously described approaches.[22,23] Thus, we investigated what types of words were most important for predicting surgery or no surgery in the following manner: (1) Example note segments were passed through the fine-tuned BERT model; (2) the self-attention mechanism sampled the query, key, and value from the given input; (3) the attention score was calculated as the matrix product of the query and key matrix; and (4) the attention score was transformed into a probability distribution with the softmax function. Finally, the attention probabilities were used as weights and combined with the value matrix to get the weighted sum of values. The higher the attention weight is, the more attention the BERT model pays to the query-key token pair and therefore the more important those tokens are to the model prediction. The words comprising the query-key token pairs at the highest (80th–90th) percentile level of attention weight in the example notes were highlighted as those words being most important to the BERT model when predicting surgery or no surgery. Additional supplemental explainability analyses to aggregate important words for prediction across the entire test set are presented in the Supplementary Materials.

Results

In this cohort, 16.6% ($n = 748$) of glaucoma patients required glaucoma surgery. Population characteristics as determined from structured data available from the EHRs are described in Table 1. Although the BERT language models could not include this structured information as input features, the information is summarized here for greater general understanding of the cohort. Mean age was 65 years old. The majority of patients were white (41.9%) and Asian (27.1%). Mean baseline IOP was 18.3 mmHg in the right eye

Table 1. Population Characteristics

	Total ($N = 4512$)	No Surgery ($N = 3764$)	Progressed to Surgery ($N = 748$)
Age (y), mean ± SD	65.0 ± 17.9	65.0 ± 18.1	64.8 ± 17.0
Right IOP, mean ± SD	18.3 ± 12.3	18.0 ± 6.2	20.1 ± 27.7
Left IOP, mean ± SD	18.8 ± 19.1	18.3 ± 6.5	21.8 ± 45.8
Right visual acuity (logMAR), mean ± SD	0.39 ± 0.74	0.39 ± 0.74	0.43 ± 0.76
Left visual acuity (logMAR), mean ± SD	0.43 ± 0.78	0.43 ± 0.79	0.43 ± 0.76
	N (%)	*N* (%)	*N* (%)
Gender (female)	2270 (50.3)	1920 (51.0)	350 (46.8)
Race			
White	1892 (41.9)	1616 (42.9)	276 (36.9)
Asian/Pacific Islander	1225 (27.1)	992 (26.4)	233 (31.1)
Other/Native American	991 (22.0)	812 (21.6)	179 (23.9)
Black	216 (4.8)	168 (4.5)	48 (6.4)
Unknown	188 (4.2)	176 (4.7)	12 (1.6)
Ethnicity			
Non-Hispanic	3791 (84.0)	3159 (83.9)	632 (84.5)
Hispanic/Latino	566 (12.5)	460 (12.2)	106 (14.2)
Unknown	155 (3.4)	145 (3.9)	10 (1.3)

Table 2. Performance Metrics for BERT Models for Predicting Glaucoma Progression to Surgery

Model	F1 (%)	Sensitivity (Recall)	Specificity	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy	Threshold
$BERT_{Base}$	0.45	0.60	0.79	0.36	0.91	0.76	0.50
$BioBERT_{v1.1+PubMed}$	0.42	0.69	0.67	0.30	0.92	0.68	0.48
$RoBERTa_{Base}$	0.45	0.40	0.92	0.50	0.88	0.83	0.83
$DistilBERT_{base}$	0.43	0.64	0.72	0.32	0.91	0.71	0.54
Clinical prediction	0.29	0.25	0.90	0.34	0.85	0.79	—

and 18.8 mmHg in the left eye. The median baseline documentation length was 1050.6 words (interquartile range [IQR], 1086) for patients who did not progress to surgery and 1159.1 words (IQR, 1043.5) for patients who progressed to surgery. Performance metrics for the different BERT models are shown in Table 2.

We developed and evaluated four different BERT-based models with regard to predicting glaucoma

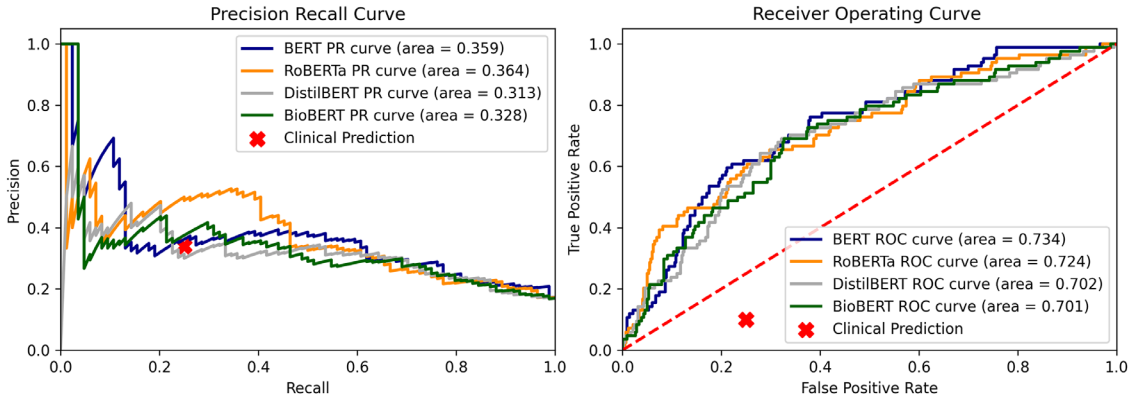

Figure 2. Receiver operating and precision recall curves for predictive models. The AUPRC (*left*) and AUROC (*right*) curves are shown for the BERT, BioBERT, RoBERTa, and DistilBERT models, as well as for the clinical prediction made by an ophthalmologist reviewing the patients' notes.

Correctly Classified
Model Prediction: Surgery 66 year old female referred for evaluation of neovascular glaucoma. History of poorly controlled diabetes and proliferative diabetic retinopathy. Has been followed locally on maximally tolerated drops, currently on latanoprost, cosopt, and brimonidine OU but IOP too high, in the 30s OD. Trial of rhopressa but had minimal effect. Has not yet tried diamox.
Model Prediction: No Surgery 47 year old man referred for evaluation of glaucoma suspect vs early POAG. Was being seen for routine optometry visit for glasses for high myopia when noted to have enlarged cup to disc ratio. Saw local ophthalmologist and was prescribed travatan qhs OU for possible visual field defect OS. Per records HVF and RNFL OCT have been stable. IOP has been well-controlled, low OU. No family history of glaucoma.
Misclassified
Model Prediction: Surgery urgent care clinic hpi- 54 y male presents to urgent care clinic today, here to establish glaucoma care. Seen by ER ophthalmology 2 days ago, found healing K ulcer OS, followed outside MD, s/p subconj ancef and vanc with hypopyon (now resolved) but found elevated IOP OS on MMT with diamox - IOP 21 by tonopen OS. S/p LPI OU. Vision improving OS but feels blurry still. Corneal ulcer evaluated given hypopyon - no posterior involvement per patient. Gtts: zioptan qhs ou, alphagan tid ou, cosopt bid ou, bacitracin-polymyxin ointment qhs os, diamox 500mg bid, besifloxacin 2 times daily os. Ocular hx: Corneal ulcer left eye followed by outside ophthalmologist without positive cultures, besifloxacin 4 times daily improving. S/p subconj ancef and vancomycin hypopyon resolved. Currently improving epi defect compared to outside ophthalmologist report. H/o hypopyon hyphema- now remaining cell. Increased IOP left eye now mmt diamox. IOP now 21 (not taking prescribed pilocarpine) but done tonopen. S/p PI both eyes. H/o diabetic retinopathy s/p focal laser. Pseudophakic both eyes. Past medical history diabetes mellitus, type 2. No family hx of glaucoma or macular degeneration.
Model Prediction: No Surgery 66 yo M with normal tension glaucoma referred for establishment of care. Pt has a history of normal tension glaucoma diagnosed approximately 10 years ago outside, with normal MRI Brain imaging. Of note, patient is of Japanese descent. Tmax of 19 OU per patient report. Has been managed on 4 medications (lumigan, timolol, dorzolamide, and alphagan) with IOP ranging from 12-14. Increasingly intolerant of medications with redness, many allergies and follicular reaction. Motivated to discontinue gtts. Has tried SLT in the past (minimal effect). Mild cataract OU. Visual fields with paracentral defects OU.

Figure 3. Example clinical notes with the most important words to model prediction highlighted. Shown are excerpts from example patient clinical notes, including those correctly and incorrectly classified by the model. Words that are most highly attended to in the model and therefore most important to the prediction are highlighted in *red*.

surgery being required. BERT had the best AUROC (73.4%), followed by RoBERTa, with an AUROC of 72.4%; DistilBERT, with an AUROC of 70.2%; and BioBERT, with an AUROC of 70.1%. Receiver operating curves and precision recall curves are shown in Figure 2. In order to qualitatively understand how the models were making their predictions, we leveraged the attention-based architecture of BERT to determine which words were most highly attended to and therefore most important for making the prediction. Example clinical notes with the most important words are provided in Figure 3. Commonly highlighted words included age, medication names, and diagnoses.

Discussion

In this study, we utilized a transfer learning strategy and fine-tuned four different BERT-based models to compare their performance with regard to predicting which subset of glaucoma patients will experience progression to require surgery based on their presenting clinical progress notes. Clinical notes contain a wealth of information related to the history of the illness, past ocular history, eye examination findings, test results, and so forth, all of which could contain key information to help predict the progression of glaucoma. Despite having different architectural characteristics and being pre-trained on different types of English-language corpora, the performance of all of the BERT models was quite similar.

Most previous predictive models for glaucoma progression have focused on using either structured data[7] or imaging and testing data, forgoing the challenge of integrating clinical free text. Baxter et al.[7] focused on using systemic data in EHRs, represented by billing codes and other similar structured variables, to predict the need for glaucoma surgery. Their most effective model, a multivariable logistic regression, reached an AUROC of 67%, a level that we were able to outperform using only free-text clinical progress

notes as inputs. Several other studies have presented methods to detect glaucoma and predict its progression with OCT using machine learning[24,25] and deep-learning techniques such as CNNs.[26] Our use of natural language processing represents a new approach to the problem of predicting glaucoma progression that focuses on integrating clinical information in free-text notes, which is different from approaches that are based on imaging and structured data and require unique data processing and modeling techniques.

Our study pioneers the investigation of using BERT-based models for ophthalmology clinical notes. We investigated BERT-based models, as they previously enabled a leap in performance on a variety of language tasks in the NLP field. BERT is a pre-trained model that can be fine-tuned to perform a variety of specific downstream tasks. Compared with training a deep-learning model from scratch, using transfer learning with pre-trained BERT models requires fewer training data. Although no previous studies have applied BERT to clinical notes to predict the progression of glaucoma, studies in other medical fields have used BERT with clinical notes to perform other clinical prediction tasks. Mohammadi et al.[14] demonstrated that BERT with clinical notes could predict the unplanned readmissions following a hip or knee arthroplasty, achieving an AUROC of 0.735. Chen et al.[13] explored the ability of BERT to identify cartilage lesions in osteoarthritis patients with radiology reports, achieving an accuracy ranging from 0.64 to 0.89, depending on the location and prevalence of the lesions. Although these previous studies were focused on different tasks in different medical fields, our study to predict glaucoma progression requiring surgery achieved a similar level of AUROC, up to 0.734 with the original BERT model, demonstrating that our BERT-based models could recognize signal in the text contributing to a glaucoma patients' prognosis. Furthermore, all of the BERT-based models we presented in this study outperformed the benchmark performance of a glaucoma specialist producing clinical predictions based on the same information on F1 score. Of course, in a true clinical assessment, glaucoma specialists would be able to directly view patient exam and test findings, so their predictions may exhibit better performance; however, until we can develop models that can also directly combine imaging features with text, the most direct comparison between model and clinician performance would be based on the free-text clinical notes alone. Our results represent an important first step toward understanding how to integrate the clinical information from clinical free text into models, suggesting that BERT-based models could be a reliable starting point.

Our study also provides important comparisons between different types of BERT models for integrating ophthalmology clinical text into predictive models. DistilBERT is the lightest model among the four BERT-based models compared, meaning that it has the fewest parameters and requires the least computational resources to train. Although the performance of the DistilBERT was 3.2% worse than BERT in AUROC score, the training time for DistilBERT was only half of the training time for BERT. It has been reported that RoBERTa outperforms BERT in General Language Understanding Evaluation benchmark results, but in our task RoBERTa demonstrated slightly weaker performance than BERT, with a 1% difference in AUROC score. Due to being pre-trained on biomedical corpora containing medical terminology and usage, BioBERT was expected to perform better than BERT in our task but did not. There are several potential reasons for this, including that both the general language corpora that BERT was pre-trained on and the biomedical literature that BioBERT was pre-trained on contain grammatically correct sentences with strong sentence-to-next-sentence relationships. However, clinical free-text progress notes in ophthalmology are often much less grammatical and may contain long lists of terms or phrases that do not have robust sentence-to-next-sentence relationships. An ophthalmology domain-specific BERT model pre-trained on ophthalmology clinical notes could mitigate these issues and potentially achieve better performance. A similar approach was taken by Mao et al.[27] to develop a domain-specific BERT related to acute kidney injury. Despite the expectation that RoBERTa should outperform BERT on general language tasks, and BioBERT should outperform BERT on biomedical domain tasks, previous studies have reported the opposite,[28] suggesting that which type of BERT model performs best may depend on the characteristics of the text data and downstream tasks.

Although BERT-based models have enabled a leap in performance in many general NLP tasks, they are not the only approach to integrating free-text notes into clinical prediction models. Neural word embeddings, in which individual words are mapped to a multidimensional vector space such that the position of each word represents its meaning,[8,29] is another approach that we have used in prior models to predict visual acuity outcomes of patients with low vision.[8] Combined with a CNN architecture designed for text, our custom ophthalmology domain-specific neural word embeddings have shown promising model prediction performance and may be a less computationally intensive method of incorporating free text into notes.[8] Thus, these and similar approaches should also continue to be

avenues of research for determining optimal methods of developing clinical prediction models based on free text.

In a similar vein, although transformer-based models are state of the art for many text-processing tasks, their staggering complexity can make them difficult to explain. Similar to previous well-accepted approaches,[23,30,31] we leveraged the attention-based mechanism of the model to examine which words were most important to the model for making predictions on specific example notes ("local explainability"). The goal was to produce explainability figures for example text that are analogous to explainability figures for imaging deep-learning models that highlight important pixels for making a prediction in individual example images. Reassuringly, our explainability studies investigating which types of words are most highly attended revealed that age, diagnoses, and use of different types of medications were important to model predictions, similar to risk factors a clinician would consider in assessing a patient's risk for surgery. It is important to note that these representations of word importance are actually one-dimensional summarizations of a two-dimensional matrix where every word and its relationship to every other word in the document have differing weights. Thus, surrounding negation and context are considered in the process by which the model attends to particular words; therefore, this type of analysis goes beyond simple counts of words or statistical comparisons between two groups (e.g., age). Although examining the attention matrix can highlight which words are important in a given example, "global" explainability methods for BERT models that attempt to characterize which words are most important for predictions in general have yet to be developed or become well established. Thus, this continues to represent an important area for future research in order to gain credibility with clinicians prior to deployment of such models in practice.

Our study has several limitations that may present opportunities for future research. First, our data and cohort were from a single academic health center. Although using transfer learning with a BERT model pre-trained on massive corpora can mitigate some challenges of training models on limited amounts of data, using larger datasets and performing external validation are ideal. Clinicians from different centers may have different patterns of writing their clinical documentation, which may impact model performance and generalizability. Sharing clinical free text containing sensitive protected health information between centers is a difficult challenge; therefore, it may be ideal to re-train or fine-tune models using data local to each health center, optimized for performance in each locale, rather than attempt to train a universally generalizable model. Second, the computational burden of BERT-based models scales quadratically with length of the input notes; thus, pre-trained models are limited to input sequences of length 512 tokens, shorter than the length of many clinical notes. To improve performance, novel methods of combining predictions on note segments, building pre-trained models that can accept longer input notes, randomly sampling words from notes, or text summarization approaches can be explored. In addition, our model was designed and trained to predict the glaucoma progression requiring surgery based on limited data from the initial presentation and was not built to update over time as new clinical information is accumulated. Future studies can investigate how best to incorporate the time dimension into these predictive models to allow for continuously updatable predictions, as well as to take into account the temporal gaps between patient encounters. Finally, BERT models do not natively incorporate structured data from EHRs; thus, future studies could develop and investigate novel deep-learning model architectures to fuse these two modalities of data, as well as compare the predictive ability of models built using these different types of data.

Conclusions

In conclusion, we present novel deep-learning approaches to predict whether glaucoma patients will progress to surgery, using transfer learning with pre-trained BERT-based transformer models on clinical free-text progress notes. Four different pre-trained BERT-based models were explored and compared, with similar results for each despite varying architectures and types of corpora used for pre-training. Our models outperformed clinical predictions by an ophthalmologist's review of the same clinical information, as well as previously published predictive models based on structured (non-text) information from electronic health records. Attention-based explainability analyses suggest that the BERT models focus on clinically relevant factors when making predictions; additional research in the field of explainable artificial intelligence may eventually develop methods to better understand these complex transformer-based deep-learning models. Nevertheless, this study is an important step toward understanding how to integrate information from clinical notes into predictive models. Future work is needed to determine how to integrate longer clinical notes, structured EHR data, and imaging data, which may further improve performance.

Acknowledgments

Supported by a grant from the National Eye Institute (1K23EY03263501 to SYW), a Career Development Award from Research to Prevent Blindness (SYW), an unrestricted departmental grant from Research to Prevent Blindness (SYW and WH), and a departmental grant from the National Eye Institute (P30-EY026877 to SYW and WH).

Disclosure: **W. Hu**, None; **S.Y. Wang**, None

References

1. Tham Y-C, Li X, Wong TY, Quigley HA, Aung T, Cheng C-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121:2081–2090.
2. Chauhan BC, Malik R, Shuba LM, Rafuse PE, Nicolela MT, Artes PH. Rates of glaucomatous visual field change in a large clinical population. *Invest Ophthalmol Vis Sci.* 2014;55:4135–4143.
3. Landers J, Martin K, Sarkies N, Bourne R, Watson P. A twenty-year follow-up study of trabeculectomy: risk factors and outcomes. *Ophthalmology*. 2012;119:694–702.
4. Abe RY, Shigueoka LS, Vasconcellos JPC, Costa VP. Primary trabeculectomy outcomes by glaucoma fellows in a tertiary hospital in Brazil. *J Glaucoma*. 2017;26:1019–1024.
5. Agrawal P, Shah P, Hu V, Khaw PT, Holder R, Sii F. ReGAE 9: baseline factors for success following augmented trabeculectomy with mitomycin C in African-Caribbean patients. *Clin Exp Ophthalmol.* 2013;41:36–42.
6. Sugimoto Y, Mochizuki H, Ohkubo S, Higashide T, Sugiyama K, Kiuchi Y. Intraocular pressure outcomes and risk factors for failure in the Collaborative Bleb-Related Infection Incidence and Treatment Study. *Ophthalmology*. 2015;122:2223–2233.
7. Baxter SL, Marks C, Kuo T-T, Ohno-Machado L, Weinreb RN. Machine learning-based predictive modeling of surgical intervention in glaucoma using systemic data from electronic health records. *Am J Ophthalmol.* 2019;208:30–40.
8. Wang S, Tseng B, Hernandez-Boussard T. Development and evaluation of novel ophthalmology domain-specific neural word embeddings to predict visual prognosis. *Int J Med Inform.* 2021;150:104464.
9. Obeid JS, Weeda ER, Matuskowitz AJ, et al. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. *BMC Med Inform Decis Mak.* 2019;19:164.
10. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. 2019, https://doi.org/10.48550/arXiv.1810.04805.
11. Sun C, Qiu X, Xu Y, Huang X. How to fine-tune BERT for text classification? *arXiv*. 2020, https://doi.org/10.48550/arXiv.1905.05583.
12. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv*. 2017, https://doi.org/10.48550/arXiv.1706.03762.
13. Chen L, Shah R, Link T, Bucknor M, Majumdar S, Pedoia V. Bert model fine-tuning for text classification in knee OA radiology reports. *Osteoarthr Cartil*. 2020;28:S315–S316.
14. Mohammadi R, Jain S, Namin AT, et al. Predicting unplanned readmissions following a hip or knee arthroplasty: retrospective observational study. *JMIR Med Inform*. 2020;8:e19761.
15. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE–an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc.* 2009;2009:391–395.
16. Wang SY, Pershing S, Tran E, Hernandez-Boussard T. Automated extraction of ophthalmic surgery outcomes from the electronic health record. *Int J Med Inform*. 2020;133:104007.
17. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics; 2020:38–45.
18. Lee J, Pershing S, Pershing S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234–1240.
19. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv*. 2019, https://doi.org/10.48550/arXiv.1907.11692.
20. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*. 2020, https://doi.org/10.48550/arXiv.1910.01108.
21. Wang SY, Hu W. Predicting glaucoma progression requiring surgery using clinical free-text notes and transfer learning with transformers. Available at: https://github.com/eyelovedata/bert_

predict_glaucoma_surgery. Accessed March 16, 2022.
22. Clark K, Khandelwal U, Levy O, Manning CD. What does BERT look at? An analysis of BERT's attention. *arXiv*. 2019, https://doi.org/10.48550/arXiv.1906.04341.
23. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv*. 2020, https://doi.org/10.48550/arXiv.1904.05342.
24. Mwanza J-C, Warren JL, Budenz DL, Ganglion Cell Analysis Study Group. Combining spectral domain optical coherence tomography structural parameters for the diagnosis of glaucoma with early visual field loss. *Invest Ophthalmol Vis Sci.* 2013;54:8393–8400.
25. Fujino Y, Murata H, Mayama C, Asaoka R. Applying "Lasso" regression to predict future visual field progression in glaucoma patients. *Invest Ophthalmol Vis Sci*. 2015;56:2334.
26. Hashimoto Y, Asaoka R, Kiwaki T, et al. Deep learning model to predict visual field in central 10° from optical coherence tomography measurement in glaucoma. *Br J Ophthalmol.* 2021;105:507–513.
27. Mao C, Yao L, Luo Y. A pre-trained clinical language model for acute kidney injury. In: *2020 IEEE International Conference on Healthcare Informatics (ICHI)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers; 2020:1–2.
28. Narayanaswamy GR. Exploiting BERT and RoBERTa to improve performance for aspect based sentiment analysis. Dublin, Ireland: Technological University Dublin; 2021. Dissertation.
29. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. Available at: https://nlp.stanford.edu/projects/glove/. Accessed March 16, 2022.
30. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18.
31. Hoover B, Strobelt H, Gehrmann S. exBERT: a visual analysis tool to explore learned representations in transformer models. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics, 2020:187–196.